Verbal and nominal components in German and Spanish phrasemes (WG 1)

Cerstin Mahlow
Insitute for Natural Language Processing (IMS)
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart
cerstin.mahlow@ims.uni-stuttgart.de

Abstract

We present first results of a systematic exploration of verbal and nominal components in German and Spanish phrasemes. The phrasemes under investigation are taken from two representative contemporary phraseological dictionaries. First results indicate similar nominal inventories for German and Spanish phrasemes but significant differences for verbs with respect to the number of verbs, their frequency, and their semantics.

1 Introduction

Our research focuses on a subset of multi-word units: phrasemes, i.e., non-Fregian discontinous multi-word expressions within sentence boundaries. Automatic processing of phrasemes poses specific challenges: (a) They usually occur within sentence boundaries, but could span two sentences, too. More often, they span several clauses within a sentence. (b) There is no fixed order of the elements of a phraseme. (c) We often find modifying adverbs or adjectives, negation, and variation in the use of definite or indefinite determiners. (d) What looks like a phraseme might be the literal use of the respective words.

Although the general linguistic assumption is that phrasemes are extremely low in frequency (Colson, 2007; Cowie, 1999), phrasemes are widely used in texts. They are of low frequency with respect to specific phrasemes—thus challenging statistic approaches for parsing or translation—but newspaper articles contain numerous phrasemes. This illustrates the importance of phrasemes and their complexity. As shown earlier (Mahlow, 2012) the creation of dedicated resources to be used in NLP is not trivial and time consuming. Therefor we should start with the most frequent elements in phrasemes, i.e., nouns and verbs occuring in a high number of phrasemes of a particular language. The research reported here aims at identifying those elements for German and Spanish.

In previous research (Juska-Bacher et al., 2013) we investigated nominal components in German and Spanish phrasemes. As material we used contemporary phraseological dictionaries for German (Dudenredaktion, 2008) and Spanish (Seco et al., 2004). First, we tested whether there is a similar list of frequent nouns in Spanish phrasemes as can be found in Russian and German phrasemes, provided by Rajchstejn (1981). We could confirm Rajchstejn's list, all nouns could be found within the 72 most frequent nouns; comparing the 50 most frequent nouns only, we found 41 corresponding nouns. For Spanish, we found 28 equivalent nouns within the 50 most frequent ones (44 nouns of Rajchstejn's list were within the 100 most frequent nouns)

Second, we compared the activity of nouns in both languages—i.e., in how many phrasemes those nouns occur—and found similarities but not a difference like Rajchstejn (1981) found for German and Russian.

2 Research question

In ongoing research, we investigate two aspects: (a) Are the verbal components used in Spanish and German phrasemes similar to the nominal components as investigated so far? (b) Can we reach general assumptions with respect to verbal components as we can do for nouns in phrasemes, i.e., can we determine equivalence relations for verbs as Hessky (1987) did for German–Hungarian nouns, Krohn (1994) for German–Swedish nouns,

and Safina (2002) for German–Russian nouns. To the best of our knowledge, there is no such in-depth research of verbal components.

3 First results

For the 14'363 German phrasemes, we found 2'973 distinct nouns and 1'604 verbs. For the 16'875 Spanish phrasemes, wie found 3'284 distinct nouns and 1'181 distinct verbs. The ratio of nouns to verbs differs considerably for German and Spanish. Figure 1 shows the activity of nouns and verbs. More Spanish nouns than German nouns occur in only one phraseme, but more German verbs than Spanish verbs occur in only one phraseme. The pattern holds for all categories.

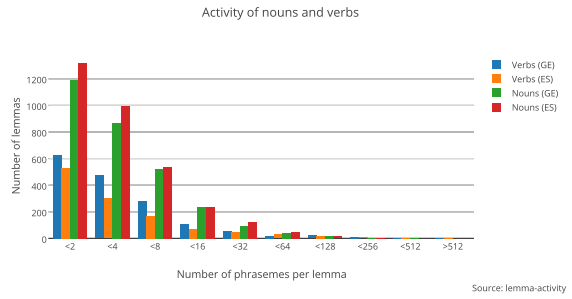

Figure 1: Activity of nouns and verbs in German and Spanish.

The most frequent 100 German nouns from the phrasemes listed in the German dictionary occur in 281 to 20 phrasemes—the most frequent noun *Hand* ('hand') occurs in 281 phrasemes, the last noun *Holz* ('wood') occurs in 20 phrasemes. The most frequent 100 Spanish nouns from the phrasemes listed in the Spanish dictonary occur in 204 to 25 phrasemes, so the distribution is similar to German. Table 1 shows the distribution of nouns with respect to semantic categories.

Semantic categories	German	Spanish
Somatisms (+ animals/humans)	38	34
Abstract words	24	46
Basic/simple elements of daily life	38	19

Table 1: Nouns

The most frequent 100 German verbs occur in 1'006 to 18 phrasemes. The most frequent 100 Spanish verbs occur in 560 to 18 phrasemes; the number of phrasemes is much smaller than for German. For verbs, we find a greater semantic variety than for nouns, shown in table 2. Distribution in German and Spanish is different.

Semantic categories	German	Spanish
Give/get/possess something	33	7
Body-related action (eat, see)	17	14
Communicate	16	7
Abstract action (love, live)	15	29
Movement	12	19
Put something somewhere	10	4
Auxiliaries/modals	8	5

Table 2: Verbs

References

Jean-Pierre Colson. 2007. The World Wide Web as a corpus for set phrases. In Harald Burger, Dmitrij Dobrovol'skij, Peter Kühn, and Neal R. Norrick, editors, *Phraseology*, Handbooks of Linguistics and Communication Science, pages 1071–1077. Walter de Gruyter, Berlin/New York.

Anthony P. Cowie. 1999. Phraseology and corpora: some implications for dictionary-making. *Lexicography*, 12(4):307–323.

Dudenredaktion. 2008. *Redewendungen: Wörterbuch der deutschen Idiomatik*. Dudenverlag, Mannheim.

Regina Hessky. 1987. *Phraseologie : linguistische Grundfragen und kontrastives Modell deutsch - ungarisch.* Niemeyer.

Britta Juska-Bacher, Cerstin Mahlow, and Sixta Quassdorf. 2013. Vergleich nominaler phraseologischer Komponenten im Deutschen und Spanischen mit computerlinguistischen Werkzeugen. In Carmen Mellado, Patricia Buján, Nely Iglesias, Mł C. Losada, and Ana Mansilla, editors, *La fraseologia del alemán y español: lexicografía y traducción*, pages 143–156. peniope, München.

Karin Krohn. 1994. *Hand und Fuss : eine kontrastive Analyse von Phraseologismen im Deutschen und Schwedischen*. Ph.D. thesis, Göteborgs Universitet.

Cerstin Mahlow. 2012. Creating a phraseme matrix based on a Tertium Comparationis. In Ruth Vatvedt Fjeld and Julie Matilde Torjusen, editors, *Proceedings of the 15th EURALEX International Congress. 7 - 11 August 2012, Oslo*, pages 720–725.

Aleksandr D. Rajchstejn. 1981. *Teksty lekcij po frazeologii sovremennogo nemeckogo jazyka voprosy frazeologiceskoj semantiki*. Pedag. Inst. Inostrannych Jazykov, Kafedra Leksikologii i Stilistiki Nemeckogo Jazyka.

Rimma Safina. 2002. Komponentenanalyse der Phraseologismen. Kontrastive Untersuchung deutsch – russisch. *Sprachwissenschaft*, 27(1):55–78.

Manuel Seco, Olimpia Andrés, and Gabino Ramos. 2004. *Diccionario Fraseologico Documentado del Español Actual: Locuciones y Modismos Españoles (Spanish Edition)*. Aguilar.